Writing Sample 3

The Feeling of the Age: A Quantitative Analysis of the Correlation Between Novelistic and Economic Sentiment

Daniella Gáti

Head note

The section you are about to read is an extract from an article currently in preparation, entitled “The Feeling of the Age: A Quantitative Analysis of the Correlation Between Novelistic and Economic Sentiment.” In this project, my coauthor and I investigate the relationship between literary and economic sentiment by correlating sentiment scores, established through text mining, with standard variables of economic sentiment such as financial indicators. The article has three main objectives. First, it aims to provide quantitative evidence for the value of fiction as data for socio-economic analysis. Thus, our project offers an argument in favor of the study of literary objects. Second, in describing a corpus of award-winning American novels since 1948 to the present using sentiment analysis, the article presents a numerically driven history of post-war American fiction. Third, the project contributes to the discussions around methods in digital humanities, in particular to the debates around the quantification of textual data.

The project is in the corpus construction stage. The extracted section comes from the beginning of the article. STRUCTURE

Is literature data? Can literary texts be mined for information about social conditions, processes and constructions? Qualitative literary study has a long tradition of analyzing literary works in order to describe social phenomena such as national founding myths, the intersection between categories of marginalization, or recurrent narratives like the white savior narrative. In these approaches, novels provide critical frameworks for a theoretical understanding of society. But do novels also speak directly to the material conditions of a particular period? In other words, do novels constitute data that is meaningful for social and economic historians?

In this article, we delineate how novels might be mined for quantitative data offering insight into the socio-economic conditions of a particular time period. We sample random extracts from 160 award-winning American novels from 1948-2018, assigning them sentiment scores which we then correlate with expressions of economic sentiment such as financial market data and structural macroeconomic indicators. We hypothesize that while novelistic sentiment may not offer predictive information about short-term economic fluctuations because of the substantial time that it takes to write and publish a novel, quantitative sentiment scores do become indicative of medium run or long run optimism or pessimism regarding the economy.

Our project contributes to multiple disciplines, especially American and literary studies, digital humanities and economics. In demonstrating quantitatively that novels constitute information directly pertaining to the conditions of society, we provide a novel, quantitative justification for the study of novels. At the same time, we complement the work of social and economic historians working in American studies by offering a quantitative description of the shifting sentiments regarding American society and the economy over the past 70 years.

In the digital humanities, a growing subfield concerns complementing traditional literary analysis by information mined computationally from literary texts. Methods like topic modeling, stylometric analysis, concordance analysis and word embedding models have been used to describe, for example, how female and male writers discuss (), or the shifting sets of words associated with characters of different genders (). Stylometric methods of authorship attribution have helped scholars establish the authorship of disputed works (). The questions that much of this scholarship seeks to answer are of a literary nature, which means that the texts have been treated as purely literary information. Our approach, however, uses text mining as a tool to interpret data on society and the economy more broadly. In so doing, we expand the stakes and significance of text mining tools in digital humanities. Thus, we respond directly to the influential critique of computational text analysis leveled by Nan Da. Da argues that quantitative literary analysis can only yield results that are either trivial or not robustly substantiated (Da), but we offer an example of results that are not trivial for either literary or economic study. By providing a quantitative analysis that adds to our understanding of the post-45 period of American literary and socio-economic history, we argue for the continued necessity of developing computational text mining methods, and suggest that the digital humanities can make meaningful contributions to fields extending outside of the humanities.

Our project is also an assessment of digital humanities methods. This assessment relates to the question of how one is to handle quantitatively the potential ambiguity of particular words and genres in relation to sentiment. Words are multivalent even before we consider the influence of context, and assigning them sentiment scores involves making an unambiguous decision about their valence in that particular context. Making the wrong decision obviously introduces noise into our econometric analysis, for which reason we conduct a range of robustness checks using variations to our sentiment analysis algorithm. Context, however, adds further complications. An essentially positive novel might be indicative of a positive outlook on society and economy, but it might just as well present an escapist fantasy of happiness precisely as succor from dire prospects. In other words, genre might be a spurious variable, causing the sentiment scores of one novel to matter potentially very differently from the scores of another. We therefore need a way to establish genre, and to use this information to systematically take genre into account when assigning sentiment scores. There already exist some sentiment analysis algorithms that take context into account to various degrees, and we provide some assessment of the respective merits of these. We experiment further with genre attribution, however, by incorporating word embedding models to determine more accurately what valence individual words carry in relation to the novel’s genre. Word embedding models, also known as word vector models, are a method of examining which words tendentially occur in similar positions, assigning high proximity scores (known as cosine similarity) to words that are used in similar ways. These models can therefore capture which words inhabit similar meanings within the document or corpus, making them a particularly powerful way to describe the specific dimensions of meaning for individual words that are active within the context of the work. As a drawback, however, they generally require larger corpora to work well. Our attempt to use them on single novels then counts as a test case for the effectiveness of such an approach when used on a smaller corpus.

Finally, our project contributes to methods of text analysis in economics, where literary texts have hitherto not been regarded as potentially rich sources of data. Computational text analysis as a tool in economics is currently gaining traction as macroeconomists attempt to use social media data to predict financial market fluctuations (Gentzkow) and microeconomists rely on text to describe phenomena that numerical analysis struggles to capture, such as discrimination in hiring (CITE). Our project complements these approaches by demonstrating how a previously entirely unused textual material, the novel, can be a rich source of economic information.

To summarize, this article makes contributions to American studies, literary studies, digital humanities methods and economics by assessing the effectiveness of text mining methods to gain non-literary information from literary texts. We quantify novelistic sentiment using sentiment analysis algorithms combined with experiments in word embedding approaches, which we then relate to economic sentiment indicators using time series and method of moments econometric analysis. Our article thus provides an evaluation of digital humanities methods, an argument for the necessity of the study of literary works, as well as a description of the shifting sentiments in post-45 American society. Before presenting our analysis, the next section considers the rationale behind our corpus and addresses the process and difficulties involved in corpus construction.

Method???

# Corpus

The corpus of literature consists of the prize-winning novels of the most prominent American literary awards for as long back as there exists information on each year’s fiction winner(s). The economic data are standard indicators of economic sentiment (interest rates, stock prices, uncertainty indexes) as well as business cycle fluctuations (GDP, inflation, unemployment) obtained from the Bureau of Economic Analysis (BEA) and the Bureau of Labor Statistics (BLS); because these data are not contentious in the economic literature, their choice will not be discussed further. The most recognized literary awards in the United States are the Pulitzer, National Book Awards, Book Critics Circle Awards, PEN/Faulkner and PEN/Hemingway awards. It will be noted that this list includes the PEN/Hemingway award for best debut novel. In sum, then, the corpus includes between one and five numbers per year in the years 1948-2018; the full list is provided in Appendix A.

The justification for relying on novels in general and on award-winners in particular has to do with several important assumptions that we make about the kind of socio-economic information that different types of literature might encode. First, award-winners in any genre can be assumed to reach a wider readership, even if we make no assumptions about whether awards do or do not reflect literary merit. In fact, because the notion of literary merit is dubious at best, it makes sense to suppose that social relevance and clarity of expression would be among the main criteria of selection for award committees. Thus, not only do such novels communicate with a relatively wide segment of society, they also reflect a judgment by literary individuals about whether important social issues are being addressed or not. Award-winning books also have the advantage that they likely minimize some of the risks associated with the kinds of escapist fantasies discussed above. Because critics on the committees of literary prizes might tendentially prioritize works that exhibit what might be called, with Matthew Arnold, “high seriousness,” the works selected by such committees are less likely to offer reading pleasure as a way to avoid dealing with unfavorable economic and social conditions, mitigating potential systematic selection bias in our data.

Second, novels, rather than poetry or drama, emerge as a desirable genre for investigating socio-economic sentiment because they have a consistently wide readership which can be presumed to inform an orientation toward questions of general social concern. It is reasonable to suppose that poetry deviates both in form and content from direct representations of sentiments about social and economic conditions, while drama does not tendentially have the same audience reach as fiction. At the same time, the post-45 American canon, which is well represented by the awards’ selections, features a significant portion of realist works that capture social and economic conditions. Well-known examples of such realist concerns among the award-winners include Harper Lee’s *To Kill a Mockingbird,* Alice Walker’s *The Color Purple,* and several of John Updike’s *Rabbit* novels. Novels, then, are arguably preferable over poetry or drama for investigating literary representations of sentiment about society at a particular time.

Postmodernist or high modernist novels complicate the situation, however. Authors like Toni Morrison, Louise Erdrich, William Faulkner, Don DeLillo or Philip Roth are represented among the award-winners with works whose rhetorical strategies foil possible expectations of the direct depiction of material conditions. This is where using sentiment analysis becomes particularly advantageous. We argue that for many of these novels, the modernist or postmodernist techniques that set them apart are achieved by evocative language use, which sentiment analysis is well equipped to capture. The haunting, mournful mood of *Beloved*, for example, or the postmodernism paranoia in *Mao II,* are arguably produced by diction that sentiment analysis will reliably recognize as negative rather than positive. Therefore, although such works do not address socio-economic conditions the way that a more realist work might, their general tone will still be recognized by sentiment analysis and interpreted in our method as indicative of assessments about the condition of society.

This discussion leads directly to the last important assumption about the literary corpus that has to be addressed: this is the notion that the underlying sentiment of a novel is related in some way to the social and economic conditions at the time of writing. An objection might be raised here that we are engaging in a type of the intentional fallacy cautioned against by Wimsatt and Beardsley (). It might be argued that we are imposing the author’s own situation as the horizon against which the sentiment of the novel becomes significant. For example, such an objection might argue that the haunting historical trauma underlying a work like *Beloved* would be misrecognized and misrepresented by an approach that seeks to attribute the tone of the novel to the race relations at the time of writing. In the case of *Beloved*, the problem can be reframed as one of presentism: our approach could be argued to disavow that historical novels might be concerned predominantly with problems in the past without necessarily transposing those problems as also relevant at the time of writing. We acknowledge these potential objections, but argue that our focus on underlying sentiment rather than on topical content justifies our approach. A historical topic, we contend, does not preclude contemporary relevance, especially when sentiments about historical processes inform the outlook that authors have about the present and future of society. We subscribe to the notion articulated by Fredric Jameson that novels embody a kind of unconscious knowledge and that this knowledge is necessarily informed by the conditions and contexts at the time of writing, as well as by expectations, desires and fears about the future (). For this reason, it makes sense to assume that literary sentiments as captured by novels have significant value as indicators of socially recognized outlooks about a society’s present and future.